DOI:
10.1111/insr.12107

Document Version
Peer reviewed version

A Review of Modern Computational Algorithms for Bayesian Optimal Design

E.G. Ryan, C.C. Drovandi, J.M. McGree, and A.N. Pettitt

School of Mathematical Sciences

Queensland University of Technology, Brisbane, Australia

email: elizabeth.ryan@hdr.qut.edu.au

Abstract

Bayesian experimental design is a fast growing area of research with many real-world applications. As computational power has increased over the years, so has the development of simulation-based design methods, which involve a number of algorithms, such as Markov chain Monte Carlo, sequential Monte Carlo and approximate Bayes methods, and which have enabled more complex design problems to be solved. The Bayesian framework provides a unified approach for incorporating prior information and/or uncertainties regarding the statistical model with a utility function which describes the experimental aims. In this paper, we provide a general overview on the concepts involved in Bayesian experimental design, and focus on describing some of the more commonly-used Bayesian utility functions and methods for their estimation, as well as a number of algorithms that are used to search over the design space to find the optimal Bayesian design. We also discuss other computational strategies for further research in Bayesian optimal design.

KEYWORDS: Bayesian optimal design; Decision theory; Utility function; Stochastic optimisation; Posterior distribution approximation.

1 Introduction

1.1 Background

Statistical experimental design provides rules for the allocation of resources in an information gathering exercise in which there is variability that is not under control of the experimenter. Experimental design has very broad applications across the natural, medical and social sciences, as well as in engineering, business and finance. Experimental designs incorporate features into studies with the aim to control systematic error (bias), reduce random variations, increase precision of parameter estimates (or some measure of interest), make predictions

about future observations, or discriminate between competing models. Essentially, non-optimal designs require more resources to make inferences on the features of interest with the same level of reward that an optimal design would. Experimental design problems are commonly viewed as optimisation problems, and optimal experimental designs may be used to achieve the experimental goals more rapidly and hence reduce experimental costs.

Experimental design has been widely developed within the classical framework, in both theory and practice (e.g., Atkinson and Donev [1992]). In the classical framework, optimal experimental designs are commonly derived using optimality criteria that are based on the expected Fisher information matrix (e.g., Fedorov [1972], Pukelsheim and Torsney [1991], Atkinson and Donev [1992]).

Classical experimental design is well suited to linear or linearised models. For nonlinear models, optimal designs generally depend on the true values of the model parameters (assuming the model is also true). Often, the aim of experimental design is to precisely estimate model parameters. Since the parameter values are not known, and data has not been collected to estimate them, the experimenter must postulate values for the model parameters from which to construct an experimental design. Use of unlikely parameter values may result in sub-optimal designs. Several studies have incorporated probability distributions on the model parameters and averaged local design criteria over the distributions so that the designs obtained may be robust to the initial choice of the parameter values (e.g., Pronzato and Walter [1985], D'Argenio [1990]). These probability distributions are known as *prior distributions* and can incorporate information from previous studies, expert elicited data or subjective beliefs of the experimenters. Similar methods are also used for situations in which there is model uncertainty (e.g., **?**). It is important to note that this prior information is subsequently ignored when performing analysis on the data generated from the experiment.

Bayesian statistics has gained popularity in the literature and has many applications, particularly in the fields of science, health and engineering. Bayesian statistics combines prior knowledge about the unknown parameters in the model with the likelihood (contribution made by the data to the unknown parameters) to give the posterior distribution, from which inferences on the unknown parameters of interest can be made.

Designs which have arisen from averaging classical design criteria over prior distributions have commonly been referred to as "Bayesian designs". We suggest this is misleading as we propose that to qualify as a "fully Bayesian design", one must obtain the design by using a design criterion that is a functional of the posterior distribution. Designs which have arisen from averaging the classical design criteria over the parameter space are termed "pseudo-Bayesian" or "robust" designs (Pronzato and Walter [1985], Federov and Hackl [1997]).

Bayesian methodologies for optimal experimental design have become more prominent in the literature (e.g., Müller [1999], Han and Chaloner [2004], Amzal et al. [2006], Müller et al. [2006], Cook et al. [2008], Huan and Marzouk [2013]). One advantage of using a Bayesian design criterion is that a single design point can be used, and the prior distribution is updated by the single observation. Lindley (1972) presents a decision theoretic

approach to experimental design, upon which Bayesian experimental design is based. Bayesian optimal design involves defining a design criterion, or a utility function $U(\mathbf{d}, \boldsymbol{\theta}, \mathbf{y})$, that describes the worth (based on the experimental aims) of choosing the design $\mathbf{d}$ from the design space $\mathbf{D}$ yielding data $\mathbf{y}$ from a sample space $\mathbf{Y}$, with model parameter values $\boldsymbol{\theta} \in \boldsymbol{\Theta}$. A probabilistic model, $p(\boldsymbol{\theta}, \mathbf{y}|\mathbf{d})$, is also required. This consists of a likelihood $p(\mathbf{y}|\mathbf{d}, \boldsymbol{\theta})$ for observing a new set of measurements $\mathbf{y}$ at the design points $\mathbf{d}$, given parameter values $\boldsymbol{\theta}$, and a prior distribution $p(\boldsymbol{\theta})$ for the parameters $\boldsymbol{\theta}$. The prior distribution is usually assumed to be independent of the design $\mathbf{d}$. A number of studies (e.g., Clyde et al. [1996], Stroud et al. [2001], Ryan et al. [2014a]) have used historical data from previous experiments to construct a prior distribution for the design of future experiments.

The Bayesian optimal design, $\mathbf{d}^*$, maximises the expected utility function $U(\mathbf{d})$ over the design space $\mathbf{D}$ with respect to the future data $\mathbf{y}$ and model parameters $\boldsymbol{\theta}$:

$$\begin{aligned}
\mathbf{d}^* &= \arg\max_{\mathbf{d}\in\mathbf{D}} \; E\{U(\mathbf{d}, \boldsymbol{\theta}, \mathbf{y})\} \\
&= \arg\max_{\mathbf{d}\in\mathbf{D}} \; \int_{\mathbf{Y}} \int_{\boldsymbol{\Theta}} U(\mathbf{d}, \boldsymbol{\theta}, \mathbf{y}) p(\boldsymbol{\theta}, \mathbf{y}|\mathbf{d}) d\boldsymbol{\theta} d\mathbf{y} \\
&= \arg\max_{\mathbf{d}\in\mathbf{D}} \; \int_{\mathbf{Y}} \int_{\boldsymbol{\Theta}} U(\mathbf{d}, \boldsymbol{\theta}, \mathbf{y}) p(\boldsymbol{\theta}|\mathbf{d}, \mathbf{y}) p(\mathbf{y}|\mathbf{d}) d\boldsymbol{\theta} d\mathbf{y}. \qquad (1)
\end{aligned}$$

Thus, the Bayesian optimal design $\mathbf{d}^*$ (given the observed data), maximises the posterior expected utility. Unless the likelihood and prior are specifically chosen to enable analytic evaluation of the integration problem, equation (1) does not usually have a closed form solution. Therefore, numerical approximations or stochastic solution methods are required to solve the maximisation and integration problem.

Practitioners have often avoided implementing Bayesian optimal design methods due to the computational difficulties involved in performing the integration and maximisation steps of equation (1). To calculate the Bayesian utility function, one must first estimate the posterior distribution (since Bayesian utilities are functionals of the posterior). Generally, one must consider thousands of posterior distributions since the posterior distribution must be calculated for each potential future data set that is drawn from the prior predictive distribution $p(\mathbf{y}|\mathbf{d}, \boldsymbol{\theta})p(\boldsymbol{\theta})$.

Bayesian design has mostly been limited to simple models (e.g., low dimensional linear and nonlinear fixed effects models). Due to the computational challenges of performing the integration and maximisation of equation (1), the use of standard optimisation algorithms, such as the Newton-Raphson method, to find the optimal design is inappropriate. This has lead to the development of novel computational strategies to solve Bayesian optimal design problems. These include: prior simulation (Müller [1999]); smoothing of Monte Carlo simulations (Müller [1999]); gridding methods which involve numerical quadrature or Laplace approximations to perform backward induction (Brockwell and Kadane [2003]); Markov chain Monte Carlo simulation in an augmented probability model (Müller [1999]); and sequential Monte Carlo methods (Kück et al. [2006], Amzal et al. [2006]). These algorithms will be discussed further in Sections 4 and 5.

1.2 Contribution and Outline

A broad range of literature exists on Bayesian optimal experimental design (e.g., Lindley [1968, 1972], Chaloner [1984], Pilz [1991], El-Krunz and Studden [1991], Müller [1999], Amzal et al. [2006]). This article aims to review those papers which reflect the computational advancements which have allowed solutions to fully Bayesian experimental design problems to be found.

Simulation-based design methods have frequently been used in the past two decades (e.g., Clyde et al. [1996], Bielza et al. [1999], Müller [1999], Stroud et al. [2001], Amzal et al. [2006], Müller et al. [2006], Cook et al. [2008], Cavagnaro et al. [2010]) in which Markov chain Monte Carlo and sequential Monte Carlo algorithms are utilised to solve complex optimal Bayesian design problems (e.g., designing for nonlinear models). Sequential, or adaptive designs, have become increasingly popular in the Bayesian design literature as they provide flexible and efficient designs. Rather than using the same design throughout the experimental process, as in *static* design problems, the design which maximises the expected utility is chosen at each stage of experimentation, based on the outcomes of previous experiments. Recent developments in static and sequential designs will be discussed further in Sections 4 and 5.

There are already several notable review papers on Bayesian experimental design. DasGupta (1995) presents a review of both classical and Bayesian experimental design, with a focus on designing for linear models. Atkinson (1996) reviews classical and pseudo-Bayesian optimal design for linear and nonlinear models. Verdinelli (1992) and Chaloner and Verdinelli (1995) present a comprehensive review on Bayesian experimental design, for both linear and nonlinear models. Müller (1999) provides an overview of simulation-based methods in optimal design. Clyde (2001) presents a broad review on several of the key concepts involved in Bayesian experimental design, such as, choice of utility functions; prior elicitation; and methods for calculating the expected utility.

There has been a lack in review papers on fully Bayesian experimental design since the early 2000s. These earlier review papers have often been written from a rather mathematical view point, and have often focused on defining Bayesian design criteria and their relationship to classical design criteria. In the past two decades there has been a substantial increase in computational power and, along with it, the use of Bayesian methodologies for optimal design. At the present time, we have been unable to find any recent review articles which discuss the various algorithms that are used in the Bayesian design literature to solve optimal design problems. Designs for complex models have also received little attention in Bayesian experimental design literature reviews. This article is concerned with reviewing the computational methods that have been used to find fully Bayesian experimental designs and aims to address the aspects of Bayesian experimental design which have received little or no emphasis in previous review papers. This article is aimed at readers with some understanding of Bayesian methods, but not necessarily with knowledge of experimental design.

In Section 2 we discuss methods for posterior distribution approximation for use in Bayesian utility functions.

In Section 3 we discuss some of the more commonly-used Bayesian utility functions, along with the methods that have been used for their estimation. Sections 4 and 5 provide an overview of the optimisation algorithms that have been used to search for static and sequential Bayesian experimental designs, respectively. We discuss future directions of Bayesian experimental design in Section 6 and provide a conclusion in Section 7.

2 Estimation of the Posterior Distribution

Bayesian utility functions are based on the posterior distribution and generally assume that a Bayesian analysis will be performed on any data that are generated from the experimental design. The utility function, when parameters are the focus of the experiment, can be a function of the scale of the posterior distribution, such as standard deviation or interquartile range. Therefore, good approximations to the posterior scale are important - not just posterior location (mean or median). In general, the posterior distribution does not have a closed form expression, and numerical methods are required to sample from or approximate the posterior distribution.

2.1 Markov Chain Monte Carlo

Markov chain Monte Carlo (MCMC) has often been used to estimate the posterior distribution for Bayesian utility function calculations (e.g., Wakefield [1994], Palmer and Müller [1998], Han and Chaloner [2004]). Although MCMC is often appropriate and useful for Bayesian data analysis, it can be too computationally intensive to perform MCMC to estimate the posterior distribution for each of the thousands of iterations required in the Bayesian experimental design algorithms (search algorithms).

2.2 Importance Sampling

Importance sampling is a popular method for estimating target distributions of interest, from which it may be difficult to sample (Geweke [1989]). Importance sampling involves choosing an *importance distribution* $g(\cdot)$, from which it is easy to sample, and then appropriately weighting the samples that have been drawn from the importance distribution to account for the discrepancy between $g(\cdot)$ and the target distribution. In the Bayesian design context, the target distribution is the posterior $p(\boldsymbol{\theta}|\mathbf{d},\mathbf{y})$. Weighted samples $\{\boldsymbol{\theta}_k, W_k\}_{k=1}^{N_p}$ are produced, where N_p is the number of particles used to estimate the posterior; $w(\boldsymbol{\theta}) = \frac{p(\mathbf{y}|\mathbf{d},\boldsymbol{\theta})p(\boldsymbol{\theta})}{g(\boldsymbol{\theta})}$ is the importance weight function; and $W_k \propto w(\boldsymbol{\theta}_k), k = 1, ..., N_p$ are the normalised importance weights, $\sum_{k=1}^{N_p} W_k = 1$. The target and importance distributions should have the same support. To measure the efficiency of importance sampling, the effective sample size (ESS) is used and can be approximated via

$$ESS = \frac{1}{\sum_{k=1}^{N_p} W_k^2}, \quad 1 \leq ESS \leq N_p.$$

Importance sampling is a very useful method for estimating the posterior distribution in Bayesian experimental design since the importance samples only need to be drawn once (unlike MCMC) and can then be re-weighted in each iteration of the optimisation algorithm according to the proposed design and data. The ability to re-use the importance samples offers substantial computational savings.

Importance sampling from the prior distribution has commonly been used in Bayesian experimental design to estimate the posterior distribution (e.g., Cook et al. [2008], McGree et al. [2012c], Ryan et al. [2014a], **?**). This reduces the importance weights to be proportional to the likelihood function. However, this is usually inefficient when there is a substantial difference between the prior and posterior distributions (e.g., Bengtsson et al. [2008], Ryan et al. [2014a], **?**).

Ryan et al. (2014a) used Laplace approximations (to the posterior) to form the importance distribution for importance sampling, and found that this approach corrects for some non-normality that is not accommodated by the Laplace approximation, and can also be used when large amounts of data are involved in the design problem since fewer particles are required in the importance sampling to obtain a reasonable ESS.

The use of adaptive importance sampling (e.g., Kinas [1996], Pennanen and Koivu [2006]) is largely unexplored for estimating the posterior distribution in Bayesian experimental design problems and may provide a fast alternative for estimating the posterior distribution. This should be considered in future research.

2.3 Deterministic Approximations

Laplace approximations (or Gaussian approximations) and numerical quadrature provide fast methods for obtaining approximations to the posterior distribution in Bayesian design problems (e.g., Lewi et al. [2009], Cavagnaro et al. [2010], Bornkamp et al. [2011], Long et al. [2013], Ryan et al. [2014a]). These methods are particularly useful when large amounts of data are involved. However, their suitability depends on whether it is reasonable to assume that the posterior distribution is well approximated by a multivariate normal distribution and they also suffer from the curse of dimensionality. To overcome the issue of dimensionality, Long et al. (2013) use polynomial-based sparse quadrature for the integration over the prior distribution.

Integrated nested Laplace approximation (INLA) is a relatively new method for rapidly approximating posterior distributions (see Rue et al. [2009]). INLA generally is a significantly faster alternative to MCMC and importance sampling for approximating the posterior. To date, INLA has mostly been used for approximate posterior inference for models in which the posterior marginals are not available in closed form due to non-Gaussian response variables, such as latent Gaussian Markov random field (GMRF) models (e.g., Rue et al. [2009]) with non-Gaussian observations. INLA enables fast Bayesian inference by using accurate approximations to the marginal posterior density for the hyperparameters and the posterior marginal densities for the latent variables. The use of INLA in the context of Bayesian experimental design is currently unexplored. For a number of

examples INLA provides good approximations to the mean and variance of the posterior distribution. Bayesian utilities depend on the posterior variance and so INLA should provide a good approximation in the context of Bayesian design. INLA also has the potential to design in the presence of random effects models, which have received little attention in the Bayesian design literature due to the difficulty of the resulting design problem.

Variational Bayesian (VB) methods facilitate approximate inference for intractable posteriors (or other target densities) and provide an alternative to other approaches for approximate Bayesian inference such as MCMC and Laplace approximations. VB methods can also be used to determine a lower bound for the evidence for use in model selection problems. The VB approach is fast and deterministic, and involves approximating the intractable target densities, e.g., $p(\boldsymbol{\theta}|\mathbf{y})$, by a factored form $q(\boldsymbol{\theta}) = q_1(\boldsymbol{\theta}_1) \times .... \times q_r(\boldsymbol{\theta}_r)$, for which $q(\boldsymbol{\theta})$ is more tractable than $p(\boldsymbol{\theta}|\mathbf{y})$. An issue is the factorization for the variational approximation $q(\cdot)$. Variational approximations have commonly been used for Bayesian inference (e.g., Ormerod and Wand [2010]), but have not yet been used in a Bayesian experimental design context. These methods could provide a fast alternative for approximating the posterior for use in Bayesian utility function calculation. However, the error of the VB approximation is generally unknown and can be substantial in terms of approximating the posterior variance (e.g., Rue et al. [2009]). Bayesian design requires a good approximation to the posterior variance, and although VB methods might approximate the mean quite well, they may not be suitable for providing good approximations to the posterior for Bayesian design.

2.4 Approximate Bayesian Computation and Other Methods for Intractable Likelihoods

Approximate Bayesian computation (ABC) is a likelihood-free method that is used to approximate the posterior distribution in situations where the likelihood function is intractable, but simulation from the likelihood is relatively straightforward. ABC has commonly been used to perform inference (e.g., Drovandi and Pettitt [2011], Drovandi et al. [2011], Sisson and Fan [2011]). One of the most common ABC algorithms is ABC rejection (see Beaumont et al. [2002]). ABC rejection prevents one from having to evaluate the likelihood by instead drawing many parameter values from the prior, and simulating data from the model, conditional on those parameter values. Only those parameters that generate simulated data that are close in some sense to the observed data are kept. The efficiency of this method is dependent on how close the posterior distribution is to the prior.

Drovandi and Pettitt (2013) and Hainy et al. (2013) used ABC rejection in the Bayesian experimental design context to approximate the posterior distributions (for Bayesian utility function calculation) for models with computationally intractable likelihoods. The ABC posterior is given by:

$$p(\boldsymbol{\theta}|\mathbf{d}, \mathbf{y}, \epsilon) = \int_{\mathbf{x}} p(\mathbf{x}|\mathbf{d}, \boldsymbol{\theta}) p(\boldsymbol{\theta}) 1(\rho(\mathbf{y}, \mathbf{x}) \leq \epsilon) d\mathbf{x},$$

where $\mathbf{y}$ represents the 'observed data' (that is generated from the model at each iteration of the optimisation (e.g., MCMC) algorithm); $\mathbf{x}$ is simulated data; $1(\cdot)$ is an indicator function; $\rho(\cdot, \cdot)$ is function that measures the discrepancy between the observed and simulated data; and ϵ is a tolerance threshold that controls the error of the approximation. The discrepancy function typically compares summary statistics of the observed and simulated data. However, Drovandi and Pettitt (2013) only considered low dimensional designs and so they were able to compare the observed and simulated data directly. ABC rejection is very useful since the ABC data, i.e., the $\mathbf{x}$ values, as well as the model parameters $\boldsymbol{\theta}$, only need to be simulated once and can be re-used at each iteration of the optimisation algorithm (much in the same spirit as importance sampling) for comparison to the observed data, $\mathbf{y}$. This offers substantial computational savings. However, the use of ABC has been limited to low dimensional designs only (i.e., up to four design points), and only discrete data has been considered.

3 Bayesian Utility Functions and Methods for Their Estimation

It is highly important that the utility function incorporates the experimental aims and is specific to the application of interest. For instance, designs which efficiently estimate the model parameters may not be useful for prediction of future outcomes. Several approaches have been suggested in the literature to assist in the elicitation of the utility function (see Spiegelhalter et al. [1996], Wolfson et al. [1996]). In practice, the utility function is often not specified as a single function, due to the difficulty of combining competing goals, and instead a set of possible utility functions is used. Christen et al. (2004) formally acknowledged the fact that the decision maker may be unwilling or unable to specify a unique utility function by considering a set of possible utility functions. Sensitivity analyses to misspecifications in the utility function have been proposed (see Rios Insua and Ruggeri [2000] for a review).

In this section we will discuss some of the more commonly used Bayesian utility functions, as well as methods for their estimation based on the approximation to the posterior. Some of the utility functions discussed in the section are the Bayesian extension to frequentist utilities, such as the alphabet criteria (e.g., A-optimality), and their connections have been outlined in Chaloner and Verdinelli (1995). One of the most commonly used and versatile Bayesian design criteria is the mutual information, which is based on entropy, and has been used for designing for efficient parameter estimation (Bernardo [1979], Ryan [2003], Paninski [2005]), as well as minimising prediction uncertainty (Liepe et al. [2013]), and model discrimination (Box and Hill [1967], Ng and Chick [2004], Cavagnaro et al. [2010], Drovandi et al. [2014]). It is a strength of Bayesian design theory that information theory provides straightforward conceptually and practically useful criteria for utility functions. There is little arbitrariness in the choice of the criteria which cover parameter estimation, data prediction and model choice.

For normal linear models, analytical expressions for equation (1) can be obtained for many Bayesian utilities, provided the model dimension and design space is small (e.g., Borth [1975], Chaloner and Verdinelli [1995], Ng

and Chick [2004]). For nonlinear design problems, one cannot usually obtain an analytical expression, and the integrals in equation (1) can instead be approximated by Monte Carlo methods (e.g., Palmer and Müller [1998], Cook et al. [2008], Ryan et al. [2014a]), Laplace approximations (e.g., Lewi et al. [2009], Ryan et al. [2014a]), or numerical quadrature (e.g., Cavagnaro et al. [2010]).

3.1 Parameter Estimation Utility Functions

Precise parameter estimation is a common goal of experimental design and many different utility functions have been used to achieve this purpose. Bayesian utility functions that design for precise parameter estimation are discussed below.

3.1.1 Information-based Utilities

When interest lies in estimating some function of $\boldsymbol{\theta}$, say $\phi(\boldsymbol{\theta})$, the mutual information between $\phi(\boldsymbol{\theta})$ and the data $\mathbf{y}$, conditional on the design $\mathbf{d}$, may be given by:

$$\begin{aligned} I(\phi(\boldsymbol{\theta}); \mathbf{y}|\mathbf{d}) &= U(\mathbf{d}) \\ &= \int_{\boldsymbol{\Theta}} \int_{\mathbf{Y}} p(\phi(\boldsymbol{\theta}), \mathbf{y}|\mathbf{d}) \Big[\log p(\phi(\boldsymbol{\theta}), \mathbf{y}|\mathbf{d}) - \log p(\mathbf{y}|\mathbf{d}) - \log p(\phi(\boldsymbol{\theta})) \Big] d\mathbf{y} d\boldsymbol{\theta} \end{aligned} \tag{2}$$

(e.g., Lindley [1956]). The optimal design that maximises the utility function is the one that yields the largest information gain, on average, about $\phi(\boldsymbol{\theta})$ upon observation of the data.

Mutual information makes use of another quantity that has also commonly been used as a Bayesian design criterion: the Kullback-Leibler divergence (KLD) (Kullback and Leibler [1951]) between the prior and posterior distributions. This is given by:

$$\begin{aligned} U(\mathbf{d}, \mathbf{y}) &= E_{\boldsymbol{\theta}|\mathbf{d},\mathbf{y}} (\log p(\phi(\boldsymbol{\theta})|\mathbf{d}, \mathbf{y}) - \log p(\phi(\boldsymbol{\theta}))) \\ &= \int_{\boldsymbol{\Theta}} p(\phi(\boldsymbol{\theta})|\mathbf{d}, \mathbf{y}) \log p(\mathbf{y}|\mathbf{d}, \phi(\boldsymbol{\theta})) d\boldsymbol{\theta} - \log p(\mathbf{y}|\mathbf{d}). \end{aligned} \tag{3}$$

Lindley (1956) suggested that this utility should be used if one is interested in maximising the expected information gain on the model parameters (or functions of) due to performing an experiment at design points $\mathbf{d}$. Mathematically, the mutual information is the KLD between the joint distribution $p(\boldsymbol{\theta}, \mathbf{y}|\mathbf{d})$ and product of marginal distributions of $\boldsymbol{\theta}$ and $\mathbf{y}$ (Borth [1975]). Alternatively, the mutual information can be thought of as the expected KLD between the prior and posterior.

Ryan (2003) used mutual information to find static designs for efficient parameter estimation. Kim et al. (2013) used the mutual information utility to find sequential designs to efficiently estimate parameters, which

was of the form:

$$U(\mathbf{d}_{(t)}) = \int_{\boldsymbol{\Theta}} \int_{\mathbf{Y}} \left[\log \left(\frac{p(\phi(\boldsymbol{\theta})|\mathbf{d}_{(t)}, \mathbf{y}_{(1:t)})}{p(\phi(\boldsymbol{\theta})|\mathbf{y}_{(1:t-1)})} \right) \right] p(\mathbf{y}_{(t)}|\mathbf{d}_{(t)}, \phi(\boldsymbol{\theta})) p(\phi(\boldsymbol{\theta})|\mathbf{y}_{(1:t-1)}) d\mathbf{y}_{(t)} d\boldsymbol{\theta},$$

where $\mathbf{y}_{(1:t)}$ are the data that were observed from the 1st to the t-th trial, $\mathbf{y}_{(t)}$ are the data that were observed at the current, t-th trial, using design $\mathbf{d}_{(t)}$, $\mathbf{y}_{(1:t-1)}$ are the data that were measured from the 1st to the $(t-1)$-th trials using the designs $\mathbf{d}_{(1:t-1)}$. Paninski (2005) proved that under acceptably weak modelling conditions, utility functions based on mutual information can choose designs that lead to consistent and efficient parameter estimates in the adaptive design framework.

Despite the theoretical appeal, mutual information is computationally complex, due to the difficulty in calculating the evidence, or marginal likelihood, $p(\mathbf{y}|\mathbf{d})$ in equation (2). Therefore, many design problems have been restricted to special cases, such as designing for parameter estimation of linear gaussian models (e.g., Lewi et al. [2009]) or binary models (e.g., Kujala and Lukka [2006]) in which the evidence can be computed analytically. Conjugate priors have been used to obtain analytic results (e.g., Borth [1975]) and numerical quadrature has also been used (e.g., Cavagnaro et al. [2010]). Drovandi et al. (2013) used sequential Monte Carlo algorithms (which are described in more detail in Sections 4 and 5) for both posterior and evidence approximation so that the mutual information could be calculated for sequential design problems for parameter estimation. Ryan et al. (2014c) used importance sampling to calculate the KLD between the prior and posterior distributions for static design problems, but found this to be computationally intensive. Huan and Marzouk (2012, 2013) used polynomial chaos approximations and nested Monte Carlo integration (Ryan [2003]) to estimate the KLD between the prior and posterior distributions for static design problems for parameter estimation.

3.1.2 Scalar Functions of the Posterior Covariance Matrix

Alternatives to utilities based on information theory are worth considering due to the general difficulty of determining the evidence $p(\mathbf{y}|\mathbf{d})$ in equations (2) and (3). Functions of the posterior distribution, such as moments, have been considered.

The inverse of the determinant of the posterior covariance matrix is a useful utility function if one is interested in maximising the (joint) posterior precision of all (or a subset) of the model parameters $\boldsymbol{\theta}$ (e.g., Drovandi et al. [2013], **?**) or a function of the model parameters $\phi(\boldsymbol{\theta})$ (e.g., Stroud et al. [2001], Drovandi et al. [2013], Ryan et al. [2014a]). This utility is also known as the "Bayesian D-posterior precision" and is given by:

$$U(\mathbf{d}, \mathbf{y}) = \frac{1}{\det(\text{cov}(\phi(\boldsymbol{\theta})|\mathbf{d}, \mathbf{y}))}.$$

If one were interested in maximising the precision of the marginal posterior distributions of the model parame-

ters, then one should use the trace instead of the determinant to obtain the Bayesian A-posterior precision. If the posterior distribution is multi-modal, then use of the Bayesian D-posterior precision utility may be inappropriate and one should instead use equation (2) as the utility function.

The posterior variance-covariance matrix can easily be obtained from the weighted posterior samples that are obtained from importance sampling (e.g., Stroud et al. [2001], McGree et al. [2012c], Drovandi et al. [2013], Ryan et al. [2014a]) and ABC rejection (e.g., Drovandi and Pettitt [2013]). The posterior variance-covariance matrix is also easily obtained when one uses numerical quadrature or Laplace approximations to the posterior distribution.

3.1.3 Quadratic Loss

When one is interested in obtaining a point estimate of the parameters, or linear combinations of them, a quadratic loss function may provide a suitable utility function:

$$U(\mathbf{d},\mathbf{y}) = -\int_{\boldsymbol{\Theta}} (\phi(\boldsymbol{\theta}) - \widehat{\phi(\boldsymbol{\theta})})^T \mathbf{A} (\phi(\boldsymbol{\theta}) - \widehat{\phi(\boldsymbol{\theta})}) p(\phi(\boldsymbol{\theta})|\mathbf{d},\mathbf{y}) d\boldsymbol{\theta},$$

where $\mathbf{A}$ is a symmetric non-negative definite matrix (e.g., Chaloner [1984], Chaloner and Verdinelli [1995], Han and Chaloner [2004]) and $\widehat{\phi(\boldsymbol{\theta})}$ is some estimate (e.g., the mean) of $p(\phi(\boldsymbol{\theta})|\mathbf{d},\mathbf{y})$. Once the posterior distribution has been approximated, it is quite straightforward to estimate this utility.

3.2 Utilities for Model Discrimination

Model discrimination is an important experimental design problem which has generated a substantial amount of research (see, for example, Box and Hill [1967], Hill et al. [1968], Borth [1975], Cavagnaro et al. [2010], Drovandi et al. [2014]). Much of the design literature has focused on producing designs that offer efficient and precise parameter estimates. However, these designs can perform poorly on model discrimination problems (see, for example Atkinson [2008], Waterhouse et al. [2009]).

Mutual information has commonly been used as the utility function in the Bayesian design literature to design for model discrimination (e.g., Box and Hill [1967], Borth [1975], Ng and Chick [2004], Cavagnaro et al. [2010], Drovandi et al. [2014], McGree et al. [2012b]). The optimal design $\mathbf{d}$ is the one that maximises the mutual information between the (random variable) model indicator, m, and the future observation $\mathbf{y}$ (see, for example, Cavagnaro et al. [2010]). Drovandi et al. (2014) give an expression of this utility to design for model discrimination for discrete data, and McGree et al. (2012b) provide an expression for continuous data. Both Drovandi et al. (2014) and McGree et al. (2012b) used sequential Monte Carlo methods to approximate the necessary quantities so that mutual information could be used to obtain sequential designs for model discrimination.

Roth (1965) proposed a model discrimination utility that is known as 'total separation', and selects design

points that yield the largest differences between the posterior predictive means of rival models. This is achieved by maximising a weighted sum (over all of the potential models) of the product of the absolute differences between the posterior predicted mean responses from all rival models and the given ('true') model. Total separation has recently been used by Masoumi et al. (2013) and McGree et al. (2012b) to design for model discrimination. The total separation utility can be approximated quite easily once the posterior predictive distribution has been found (see, for example McGree et al. [2012b]). This utility does not account for the variance of the predicted responses (Hill [1978]), which is problematic if the competing models differ in their error structures (e.g., additive vs. multiplicative error) (McGree et al. [2012b]).

Both mutual information and total separation do not rely on the assumption of a particular model being true (unlike many of the classical design criteria), but require the experimenter to define a set of rival models with prior probability of being true. That is, these utilities use the M-closed approach of Bernardo and Smith (2000, chapter 6).

Vanlier et al. (2014) proposed a model discrimination utility that is based on a k-nearest neighbour estimate of the Jensen Shannon divergence (which is the averaged KLD between the probability densities and their mixture) between the multivariate predictive densities of competing models. They showed that their utility is monotonically related to the expected change in the Bayes Factor in favour of the model that generated the data. MCMC was used to sample from the posterior distributions and the predictive distributions were sampled using these posterior distribution values and by adding noise generated by the error model. This was found to be computationally intensive, especially for their application which involved nonlinear models of biochemical reaction networks.

3.3 Utilities for Prediction of Future Observations

If one is interested in choosing $\mathbf{d}$ to predict $\mathbf{y}_{n+1}$ from $\mathbf{y} = (\mathbf{y}_1, ..., \mathbf{y}_n)$, then the expected gain in Shannon information (Shannon [1948]) (or the expected KLD) for a future observation, $\mathbf{y}_{n+1}$, from the prior predictive distribution to the posterior predictive distribution can be used as the utility function:

$$U(\mathbf{d}_{(n+1)}, \mathbf{y}) = \int_{\boldsymbol{\Theta}} \int_{\mathbf{Y}_{n+1}} p(\mathbf{y}_{n+1}|\mathbf{d}_{n+1}, \mathbf{y}_{1:n}, \phi(\boldsymbol{\theta})) \log p(\mathbf{y}_{n+1}|\mathbf{d}_{n+1}, \mathbf{y}_{1:n}, \phi(\boldsymbol{\theta})) d\mathbf{y}_{n+1} d\boldsymbol{\theta} - \log p(\mathbf{y}_{1:n}|\mathbf{d}_{1:n}),$$

(e.g., Chaloner and Verdinelli [1995] and references therein). This is equivalent to the mutual information between the future observation $\mathbf{y}_{n+1}$ and the previous observations $\mathbf{y}_{1:n}$, conditional on the future designs $\mathbf{d}_{n+1}$ and previous designs $\mathbf{d}_{1:n}$. Leipe et al. (2013) used mutual information to minimise prediction uncertainty in sequential systems biology experiments. Zidek et al. (2000) used maximum entropy to obtain designs that maximised information about expected responses for air quality monitoring sites.

Geostatistical design problems often use utilities that are functions of the prediction variance. For example, Diggle and Lophaven (2006) propose a Bayesian design criterion that chooses a set of sampling locations to enable efficient spatial prediction by minimising the expectation of the spatially averaged prediction variance (with respect to the marginal distribution of the data).

If one is interested in minimising the variance of the expected response, then one could use the utility function developed by Solonen et al. (2012) which places the next design point where the prior variance of the mean response is largest. The utility is calculated by bringing in the observations one-at-a-time and is given by:

$$U(\mathbf{d}, \mathbf{y}) = \prod_{k=1}^{K} (\sigma^2 + \text{Var}_{\boldsymbol{\theta}|\mathbf{y}_{1:(k-1)}}(m_k(\boldsymbol{\theta}))), \tag{4}$$

where σ^2 is the residual variance, $m_k(\boldsymbol{\theta}) = E(y_k|d_k, \boldsymbol{\theta})$ and K is the number of observations.
The expression $\text{Var}_{\boldsymbol{\theta}|\mathbf{y}_{1:(k-1)}}(m_k(\boldsymbol{\theta}))$ gives the variance of the mean response at d_k, given measurements $\mathbf{y}_{1:(k-1)}$ at points $\mathbf{d}_{1:(k-1)}$. The utility at d_k is evaluated using a weighted variance, where each simulated response is weighted based on the likelihood of previous simulated measurements, $p(\mathbf{y}_{1:(k-1)}|\mathbf{d}_{1:(k-1)}, \boldsymbol{\theta})$.

Solonen et al. (2012) advocate the use of this utility function to design for parameter estimation since it is easier to compute than information-based utility functions (equation (2)) since it does not require evidence calculation. Solonen et al.'s (2012) utility function assumes a constant variance. Ryan et al. (2014c) present a generalised version of this utility function which may be used when the error structure of a model has a non-constant variance. Ryan et al. (2014c) found that Solonen et al.'s (2012) utility function did not perform well when designing for parameter estimation.

3.4 Utilities for Several Design Objectives

Researchers often have several competing goals for an experiment, rather than one single goal, and so these competing design objectives can be incorporated into one or several utility functions. One approach to dealing with competing design objectives is to weight each design criterion and search for the design that optimises the weighted average of these criteria. This is known as a compound or weighted design problem (e.g., Dette [1990]). Clyde and Chaloner (1996) discuss compound design criteria and present an equivalence theorem for Bayesian constrained design problems. DasGupta et al. (1992) gave examples of compromise designs in which one is interested in finding a design that is highly efficient for several design problems.

Borth (1975) extends the mutual information utility proposed by Box and Hill (1967) so that fully Bayesian designs could be obtained for the dual goals of model discrimination and parameter estimation. This utility is known as "Total entropy". This dual design problem has been investigated in a number of classical design papers through use of compound criteria such as $D|T$- and $T|D-$optimality and hybrid DT-optimality (e.g., Atkinson [2008], Tommasi [2009], Waterhouse et al. [2009]), but is largely unexplored in the Bayesian design literature.

Chaloner and Verdinelli (1995) discuss several Bayesian utility functions that may be used for the dual purpose of maximising the expected value of the response and the expected information gain, and utilities which may be used to design for parameter estimation and prediction.

McGree et al. (2012c) considered compound utility functions in the context of Bayesian adaptive designs for dose-finding studies for the dual design objectives of estimating the maximum tolerated dose and addressing the safety of the study subjects. A number of different estimation utilities were used, and the utility functions only allowed doses to be available for selection if the 95th percentile of the posterior predictive probability of toxicity was less than some pre-specified tolerance level. Drovandi et al. (2013) developed a hybrid utility function for an adaptive dose-finding study to obtain robust estimates of the target stimulus-response curve in the presence of model and parameter uncertainty.

A number of studies have had the dual objectives of designing for parameter estimation or prediction accuracy and to minimise study costs (or inconvenience to study subjects). Stroud et al. (2001) used utility functions which designed for the precise estimation of parameters of interest, as well as minimising inconvenience to study subjects by penalising samples that were collected after a certain time period. Palmer and Müller (1998) searched for the optimal sampling times for stem cell collections in cancer patients, to minimise the expected loss function over the posterior predictive distribution for a new patient. Their utility function also included a penalty for failing to collect a certain target number of stem cells and a cost penalty for each sampling time scheduled.

4 Static Design Search Algorithms

Now that we have described methods for estimating $U(\mathbf{d}, \mathbf{y})$, we will now discuss the algorithms in which they are embedded to calculate and maximise $U(\mathbf{d})$.

Static design problems assume that the same design will be used throughout the experimental process, regardless of the incoming information that may be collected from the experiment. Static designs are useful when data are collected in a batch, according to a fixed protocol. Static designs are also useful for experiments in which data are not available until a considerable time after treatment allocation. A number of different algorithms have been used to solve Bayesian static design problems and they will be discussed below. These include: prior simulation (Müller [1999]); smoothing of Monte Carlo simulations (Müller [1999]); MCMC simulation in an augmented probability model (Müller [1999]); and sequential Monte Carlo (SMC) methods (Kück et al. [2006]).

4.1 Monte Carlo Integration

In many situations, one can simulate values of $(\boldsymbol{\theta}_i, \mathbf{y}_i)$ (for $i = 1, ..., M$) from $p(\boldsymbol{\theta}, \mathbf{y}|\mathbf{d})$ and the utility function can be estimated using these values. The integral is approximated by using:

$$\hat{U}(\mathbf{d}) = \frac{1}{M} \sum_{i=1}^{M} U(\mathbf{d}, \boldsymbol{\theta}_i, \mathbf{y}_i). \tag{5}$$

The optimal design, $\mathbf{d}^* = \arg\max \hat{U}(\mathbf{d})$, can then be found by using a suitable maximisation method to search over the estimates, $\hat{U}(\mathbf{d})$ (see Müller [1999]). This approach has commonly been used in the literature (e.g., Wakefield [1994], Carlin et al. [1998], Palmer and Müller [1998]) and is useful when a discrete set of possible designs that are of low dimension are used.

Müller and Parmigiani (1995) use a similar approach to equation (5), in which stochastic optimisation is performed by fitting curves to the Monte Carlo samples. First, they simulate draws from $(\boldsymbol{\theta}, \mathbf{y})$ and evaluate the observed utilities. Then, a smooth curve is fitted through these simulated points, which serves as an estimate of the expected utility surface. The optimal design can then be found deterministically. Kuo et al. (1999) also used these curve fitting methods for solving design problems of low dimension.

Straightforward Monte Carlo integration over $(\boldsymbol{\theta}, \mathbf{y})$ for each design $\mathbf{d}$ may be computationally intensive for design problems involving a large number of design variables, since the design space grows far too rapidly with the number of design variables and thus the grid search over the design space becomes infeasible. Also, when a design variable corresponds to a data point, then a larger number of design variables means that more observations are involved, which implies a larger integral over $\mathbf{y}$, and thus a larger value of M is required to accurately estimate $U(\mathbf{d})$. Therefore, alternative methods are often required to solve the optimisation problem.

4.2 MCMC Algorithms

4.2.1 MCMC Simulation in an Augmented Probability Model

Alternatively, Clyde et al. (1996), Bielza et al. (1999) and Müller (1999) solved the optimal design problem by treating the expected utility as an unnormalised marginal probability density function. This was achieved by placing a joint distribution on the target function to form an augmented probability model, which is given by:

$$h_J(\mathbf{d}, \boldsymbol{\theta}_1, ..., \boldsymbol{\theta}_J, \mathbf{y}_1, ..., \mathbf{y}_J) \propto \prod_{j=1}^{J} U(\mathbf{d}, \boldsymbol{\theta}_j, \mathbf{y}_j) p(\boldsymbol{\theta}_j, \mathbf{y}_j|\mathbf{d}), \tag{6}$$

where J is a fixed (and usually large, say 20 or higher) integer. For each $\mathbf{d}$, one simulates J experiments $(\boldsymbol{\theta}_j, \mathbf{y}_j), j = 1, ..., J$, independently from $p(\boldsymbol{\theta}, \mathbf{y}|\mathbf{d})$ and considers the product of the calculated utilities. The product of the calculated utilities (rather than the sum) is used to ensure that the marginal distribution in $\mathbf{d}$

is proportional to the expected utility, i.e., $h_J(\mathbf{d}) \propto U^J(\mathbf{d})$. It is assumed that $U(\mathbf{d}, \boldsymbol{\theta}_{1:J}, \mathbf{y}_{1:J})$ satisfies the appropriate conditions for $h_J(\cdot)$ to be positive and integrable over $(\mathbf{D}, \boldsymbol{\Theta}, \mathbf{Y})$. The design space $\mathbf{D}$ is assumed to be bounded.

One can then use a Metropolis-Hastings (MH) MCMC scheme to simulate from $h_J(\cdot)$ and select random draws from the design space that are proportional to the utility that is attached to the design. The MH MCMC algorithm focuses on sampling designs in areas of high expected utility and discourages sampling in areas of low expected utility (see Müller [1999]). The sample of simulated $\mathbf{d}$ may be used to provide an estimate of $h_J(\mathbf{d})$ and the joint mode of $h_J(\mathbf{d})$, $\mathbf{d}^*$, corresponds to the optimal design. Algorithm 1 describes the process involved to simulate from $h_J(\cdot)$.

Algorithm 1: MCMC algorithm for Bayesian optimal design

1 Initialise - set an initial design $\mathbf{d}^{(1)}$, simulate $(\boldsymbol{\theta}_j, \mathbf{y}_j)$ from $p(\boldsymbol{\theta}, \mathbf{y}|\mathbf{d}^{(1)}) = p(\boldsymbol{\theta})p(\mathbf{y}|\mathbf{d}^{(1)}, \boldsymbol{\theta})$ for $j = 1, ..., J$

2 Compute $U^{(1)} = \prod_{j=1}^{J} U(\mathbf{d}^{(1)}, \boldsymbol{\theta}_j, \mathbf{y}_j)$

3 **for** *i = 1 to iters* **do**

4 Generate a candidate design $\tilde{\mathbf{d}} \sim q(\cdot|\mathbf{d}^i)$, propose $(\tilde{\boldsymbol{\theta}}_j, \tilde{\mathbf{y}}_j) \sim p(\boldsymbol{\theta}, \mathbf{y}|\tilde{\mathbf{d}}) = p(\boldsymbol{\theta})p(\mathbf{y}|\tilde{\mathbf{d}}, \boldsymbol{\theta})$ for $j = 1, ..., J$

5 If $\tilde{\mathbf{d}}$ is not within the design space then reject the proposal and go to line 9

6 Compute $\tilde{U} = \prod_{j=1}^{J} U(\tilde{\mathbf{d}}, \tilde{\boldsymbol{\theta}}_j, \tilde{\mathbf{y}}_j)$

7 Calculate the MH acceptance probability, $a = \min(1, A)$ where

$$A = \frac{\tilde{U} \times q(\mathbf{d}^{(i)}|\tilde{\mathbf{d}})}{U^{(i)} \times q(\tilde{\mathbf{d}}|\mathbf{d}^{(i)})}$$

Here $U^{(i)}$ and $\mathbf{d}^{(i)}$ are the current utility and design point values, respectively, and $\tilde{U}$ and $\tilde{\mathbf{d}}$ are the proposed utility and design point values, respectively.

8 Set $(\mathbf{d}^{(i+1)}, U^{(i+1)}) = (\tilde{\mathbf{d}}, \tilde{U})$ with probability a, and

9 $(\mathbf{d}^{(i+1)}, U^{(i+1)}) = (\mathbf{d}^{(i)}, U^{(i)})$ with probability $1 - a$.

We note that the joint mode of $h_J(\mathbf{d})$ needs to be found rather than the marginal modes for each element of $\mathbf{d}$ as the latter may be very different from the former. Cook et al. (2008) and Drovandi and Pettitt (2014) propose methods for searching for the multivariate mode using a non-parametric density estimate of the (annealed) expected utility surface based on the design samples obtained from the MCMC. However, for design problems that involve a moderate number of design points ($dim(\mathbf{d}) \geq 4$), the problem of finding the multivariate mode is more difficult than finding marginal modes and one may need to use dimension reduction techniques, such as those that Ryan et al. (2014c) propose which project the design space onto a lower dimensional space. However, dimension reduction techniques may not always be appropriate and further research is needed into the sub-optimality of finding the multivariate mode for a large number of design variables.

In some design problems the range of values taken by the utility in the neighbourhood of its mode can be sufficiently small so that the Monte Carlo error can dominate this range. Then the mode is difficult to locate accurately. However, the problem can be mitigated by the fact that there is a neighbourhood of designs with

near optimum utility and exact location of the mode is therefore not necessary.

4.2.2 Simulated Annealing-type Approach

Müller (1999) proposes an extension to "MCMC Simulation in an Augmented Probability Model" in which the J values are increased to make the expected utility surface more peaked. This does not change the solution of the optimal design problem. This approach has been very popular in the literature (e.g., Müller [1999], Stroud et al. [2001], Cook et al. [2008]) and uses similar ideas to simulated annealing (see Van Laarhoven and Aarts [1987]) where $T = 1/J$ may be interpreted as the "annealing temperature". As $T \to 0$, the original target function is replaced with a point mass at the mode (Müller [1999]). As J increases, the utility surface will become more peaked and simulations will cluster more tightly around the mode. However, increasing J obviously increases the number of required computations. An annealing schedule is not necessarily required, i.e., the same value of J may be used for all simulations. However, this is not efficient for high dimensional problems (see Amzal et al. [2006]) and a "cooling" schedule may be required where J increases to $+\infty$. Müller et al. (2004) recommend that J should be gradually increased as the algorithm progresses so that the search will not become trapped in a local mode for situations where several modes exist. In Müller et al.'s (2004) approach, the algorithm initially explores the entire design space, but as the J value increases, the MCMC draws focus around one of the highest modes. One can embed Algorithm 1 into an annealing schedule to increase J over the iterations.

Whilst the algorithm presented by Müller (1999) has "theoretically appealing" properties (i.e., one can sample from the expected utility surface using a MH MCMC algorithm in which sampling is focused in areas of a high expected utility; and as $J \to \infty$, the expected utility is replaced with a point mass at the mode), it has been found to have slow convergence in practice, particularly for situations where there are a large number of design variables for which this algorithm becomes inefficient (Stroud et al. [2001], Amzal et al. [2006]). Use of this algorithm has therefore mostly been restricted to up to four design variables (e.g., Bielza et al. [1999], Müller [1999], Stroud et al. [2001], Cook et al. [2008]) and further research is required for searching for solutions to high dimensional design problems.

4.3 SMC Algorithms

Should I put in an algorithm like the one in James' annealing paper? If so, James, can you send me the pseudo code from latex to save me typing it out, please?

SMC algorithms, also known as particle filters, use a population of particles to approximate a distribution and move through a smooth sequence of connected target distributions using resampling and diversification of particles until the final target distribution is reached (see Chopin [2002], Del Moral et al. [2006]). SMC combined with Markov and MCMC kernels provides a powerful and efficient computational approach for approximating

target distributions. SMC has only been applied to static design problems in a few instances (see Amzal et al. [2006], Kück et al. [2006]).

SMC methods can be useful for sampling from target distributions that change. This also includes the target distribution $h_J(\mathbf{d}, \boldsymbol{\theta}_{1:J}, \mathbf{y}_{1:J})$ (Müller et al. [2004]) in which J increases. For nonlinear and high dimensional design problems, Amzal et al. (2006) extend the approach of Müller (1999) and Müller et al. (2004) by using SMC methods to build a sequence of target distributions that were based on the annealed $h_J(\cdot)$. At iteration $t-1$, the particle set $\{\mathbf{d}_k^{(t-1)}, W_k^{(t-1)}\}_{k=1}^{N_p}$ (where N_p is the number of particles) provides an approximation for $h_{J(t-1)}$. A re-weight step is then implemented in the SMC algorithm via importance sampling to update the weighted particle set to approximate $h_{J(t)}$. Particles with a higher utility are given more weight than those with a lower utility. As J increases, the target distribution becomes more peaked around the mode. Resampling and mutation steps are also used to avoid denegeracy in the particle set.

Kück et al. (2006) use SMC methods to generalise the approach of Müller et al. (2004) to non integer annealing steps. Kück et al.'s (2006) approach was found to behave well when exploring multi-modal target distributions. The choice of how to increase $J(t)$, where t is the iteration number, is important since large increments can result in degeneracy of the particles and small increments are computationally inefficient. McGree et al. (2012a) propose to choose the increment to maintain a specific level of efficiency (based on the ESS) in the sample.

4.4 Other Stochastic Approximation Algorithms

Huan and Marzouk (2013) used simultaneous perturbation stochastic approximation (SPSA) (Spall [1998]) and Nelder-Mead nonlinear simplex (NMNS) (Nelder and Mead [1965]) algorithms to perform stochastic optimisation for nonlinear and computationally intensive models. These algorithms were used to maximise expected utility functions that were estimated via Monte Carlo integration. Polynomial chaos surrogate models were used to simulate data from the computationally intensive models.

SPSA is a stochastic approximation method that is similar in nature to a steepest-descent method that uses a finite difference estimate of the gradient. However, SPSA only uses two random perturbations to estimate the gradient, regardless of the dimension of the problem. Whilst the finite differences stochastic approximation (FDSA) algorithm only perturbs in one direction at a time, the SPSA algorithm perturbs in all directions at once. In SPSA, the error in the estimation of the gradient is "averaged out" over a large number of iterations (Spall [1998]) and the algorithm has a similar convergence rate to FDSA. SPSA has a global convergence property that relies on the existence of a non-negligable noise level in the objective function and the finite-difference-like perturbations (Maryak and Chin [2004]). However, high noise levels can cause slow convergence or can cause the algorithm to become stuck in local optima. SPSA is suitable for large-scale population models.

The NMNS algorithm has commonly been used for deterministic optimisation of nonlinear functions. It is

a well-studied numerical method that is useful for problems in which gradients may be unknown. The NMNS algorithm is useful when dealing with noisy objective functions since it only requires a relative ordering of the function values, rather than the magnitudes of the differences (as when estimating gradients). NMNS is less sensitive than SPSA to the noise level, but can converge to non-optimal points. Huan and Marzouk (2013) found that the NMNS algorithm performed better than SPSA overall, in terms of the asymptotic distribution of the design variables and how quickly convergence was achieved.

Huan and Marzouk (2012) used the Robbins-Monro (RM) (Robbins and Monro [1951]) stochastic approximation, and compared it to a sample average approximation combined with the Broyden-Fletcher-Goldfarb-Shanno method (SAA-BFGS) to solve the optimal design problem for models that were described by the solution to partial differential equations. Polynomial chaos surrogates were used to simulate from the model.

The RM algorithm is one of the oldest stochastic approximation methods. It uses an iterative update that is similar to steepest descent, but uses stochastic gradient information. Sampling average approximation (SAA) algorithms reduce a stochastic optimisation problem to a deterministic one. For instance, in the optimal experimental design framework, we may define the problem to be solved as:

$$\mathbf{d}^* = \arg\max_{\mathbf{d}\in\mathbf{D}}\{U(\mathbf{d})\} = \arg\max_{\mathbf{d}\in\mathbf{D}} E_W[\hat{U}(\mathbf{d}, W)],$$

where $\mathbf{d}$ is the design variable, W is the "noise" random variable, and $\hat{U}(\mathbf{d}, W)$ is an unbiased estimate of the objective function, $U(\mathbf{d})$ (e.g., the expected KLD between the prior and posterior distributions). SAA approximates this optimisation problem using

$$\hat{\mathbf{d}}_s = \arg\max_{\mathbf{d}\in\mathbf{D}}\{\hat{U}_M(\mathbf{d}, w_s) \equiv \frac{1}{M}\sum_{i=1}^{M}\hat{U}(\mathbf{d}, w_i)\},$$

where $\hat{\mathbf{d}}_s$ and $\hat{U}_M(\mathbf{d}, w_s)$ are the optimal design and utility function values under a particular set of M realisations of W, where $w_s \equiv \{w_i\}_{i=1}^M$. The same set of realisation of W is used for different values of $\mathbf{d}$ throughout the optimisation process, which makes the maximisation problem deterministic. The Broyden-Fletcher-Goldfarb-Shanno (BFGS) method (Nocedal and Wright [2006]), which is a deterministic quasi-Newton method, was used to find $\hat{\mathbf{d}}_s$ as an approximation to $\mathbf{d}^*$.

Huan and Marzouk (2012) used infinitesimal perturbation analysis (Ho and Cao [1983]) to construct an unbiased estimator of the gradient of the KLD for use in the RM algorithm. A polynomial chaos approximation of the forward model was also used to speed up computation of the utility function and gradient evaluations. Huan and Marzouk (2012) found that, although SAA-BFGS generally required fewer iterations, each iteration had a longer run time than a step of RM. As the evaluation of the utility function becomes more expensive, RM may be the more suitable of the two methods. RM was also found to outperform SAA-BFGS in terms of the

size of the mean square error (between the "true" optimal value of the KLD and the value of the KLD for the current iteration), for a given computational effort.

5 Sequential Design Search Algorithms

Decisions are often made in stages, with additional data being observed between the decisions. For example, in dose-finding trials, dose allocation decisions are often made after previous cohorts have been administered the treatment so that future cohorts may be given doses that are closer to the maximum tolerated dose. Whitehead and Brunier (1995) and Whitehead and Williamson (1998) implement a Bayesian m-step look-ahead procedure to find the optimal treatment dose to administer to the next m patients in a dose-finding study. Sequential design problems are those that involve an alternating sequence of decisions and observations. The Bayesian paradigm is extremely useful for sequential design problems since the posterior can be used as the prior distribution for the next experiment.

5.1 Backwards Induction

Although many approaches to solving sequential design problems use a myopic approach, which involves looking ahead only to the next observation (e.g., Cavagnaro et al. [2010], Drovandi et al. [2014], McGree et al. [2012b]), in general, this is not optimal, and one should instead look ahead to all future observations in the experiment (Borth [1975]), as well as the decisions that might be made at each future observation. To achieve this, the computationally intensive *backward induction* method should be used (see, for example, DeGroot [1970], Berger [1985], Bernardo and Smith [2000] for a description) which considers all future observations. Backward induction is also known as stochastic dynamic programming (e.g., Ross [1983]).

Early work in this area was restricted to simple model settings, such as one-sided tests of a univariate parameter (Berry and Ho [1988]), and binary outcome settings (Lewis and Berry [1994]). These approaches typically used only two or three backwards steps (interim looks at the data). Carlin et al. (1998) extend these approaches by including a forward sampling algorithm that can be used to find the optimal stopping boundaries in clinical trials and eases the computational burdens associated with backward induction. However, Carlin et al. (1998) used a univariate normal likelihood, assumed that the standard deviations were known at each step, and considered a maximum of 4 backwards steps.

Brockwell and Kadane (2003) proposed a gridding method which approximates the expected loss function (utility function) at each decision time, and consists of a function of certain summary statistics (low dimensional) of the posterior distribution of the parameter of interest. Their approach is similar to that of Berry et al. (2000). Brockwell and Kadane (2003) use a one-step-ahead forward simulation procedure to evaluate the expected utilities and focus on problems related to parameter estimation. Müller et al. (2006) also use a similar approach

to Brockwell and Kadane (2003) which involves forward simulation to approximate the utility functions and constrain the action space to circumvent the problem of an increasing number of possible trajectories in the backward induction steps. Rossell et al. (2007) extend the approaches of Carlin et al. (1998), Brockwell and Kadane (2003), and Müller et al. (2006), in which they compute a summary statistic when new data are observed and use decision boundaries that partition the sample space. Once the summary statistic falls in the stopping region, the experiment is terminated. Thus the sequential problem is reduced to the problem of finding optimal stopping boundaries, and the choice of these boundaries accounts for all future data. Rossell and Müller (2013) extend these ideas to high dimensional data by assuming that the data are suitably pre-processed.

Backwards induction is still limited to simple design problems, such as stop/continue decisions in dose-finding trials.

5.2 MCMC Algorithms

Put in algorithm? If so, James, do you have a general version of algorithm 2 from your paper that I could put in?

McGree et al. (2012c) used MCMC methods (MH algorithms) to sample from the posterior distribution to find adaptive designs for a dose-finding study. Bayesian compound utility functions were used to find the dose for the next subject for the dual purposes of estimating the maximum tolerated dose (MTD) and addressing safety issues of toxicity. To estimate the utility functions, importance sampling was used in which the posterior distribution of the parameters (using the observations up to the *i-1*th subject) $p(\boldsymbol{\theta}|\mathbf{y}_{(1:i-1)})$ was used as the importance distribution, and the target distribution was $p(\boldsymbol{\theta}|\mathbf{y}_{(1:i)})$, where $\mathbf{y}_i$ is the new data point given by dose D. McGree et al.'s (2012c) algorithm involved a form of self-tuning in that the proposal distribution for the model parameters $\boldsymbol{\theta}$ was based on a bivariate normal distribution in which the mean and variance were obtained from a maximum likelihood fit to the current data. Each time a new dose was selected, the proposal distribution was updated. However, re-running MCMC after each observation is taken is a very computationally expensive process.

5.3 SMC Algorithms

Should I put an algorithm or 2 in here? Given the ones for parameter estimation are different to model uncertainty, so I'm not sure what to put in....

SMC improves upon the MCMC approach for sequential design problems since new observations can be included via a simple re-weighting approach and can be helpful for estimating utilities, such as the mutual information, since SMC produces an estimate of the evidence as a by-product. SMC has been used for parameter estimation design problems (e.g., Drovandi et al. [2013]), and model discrimination design problems (see Cav-

agnaro et al. [2010], Drovandi et al. [2014]). Its design applications are diverse and include computer experiments (e.g., Loeppky et al. [2010]), astrophysics (e.g., Loredo [2004]), cognitive science (e.g., Cavagnaro et al. [2010]), neurophysiology experiments (e.g., Lewi et al. [2009]), clinical trials (e.g., Liu et al. [2009]) and bioassays (e.g., Tian and Wang [2009]).

Cavagnaro et al. (2010) use a similar approach to Amzal et al. (2006) in which an SMC algorithm was implemented to design optimally for model discrimination in the context of memory retention models. A simulated annealing effect (Müller [1999]) was used in which the utility function was incrementally "powered up". Cavagnaro et al.'s (2010) SMC algorithm designs for experiments one-observation-at-a-time, using the posterior distribution that is based on all of the data that has been observed thus far. Whilst these myopic approaches are sub-optimal, they are necessary in many applications of Bayesian design of experiments due to computational complexity of the backwards induction algorithm (Section 2.6.1).

Drovandi et al. (2014) present an SMC algorithm to sequentially design experiments one-at-a-time in the presence of model uncertainty for discrete data. McGree et al. (2012b) extended this approach for continuous data. In these works, an SMC algorithm is run in parallel for each of the competing models and the results are combined to compute the utility function in the presence of model uncertainty. This algorithm avoids between-model or cross dimensional proposals. The SMC algorithm produces an approximation to the evidence as a by-product (Del Moral et al. [2006]), which is used to compute the posterior model probabilities and to estimate the utility function. This avoids the need to use computationally intensive techniques, such as quadrature (e.g., Cavagnaro et al. [2010]) to obtain an estimate of the evidence. Once the posterior model probabilities are computed, model discrimination utility functions (see Section 3.2), that are derived from information theory, such as the entropy of model probabilities (Box and Hill [1967], Borth [1975]) can be evaluated. The design d that is chosen is the one that maximises the mutual information between the model indicator, m, and the predicted observation (Cavagnaro et al. [2010]). Little problem specific tuning is required for this algorithm and it is much less computationally intensive than approaches that rely on MCMC for posterior simulation in sequential design contexts (e.g., McGree et al. [2012c], Section 2.6.2).

In both Drovandi et al. (2014) and McGree et al.'s (2012b) work, only a discrete design space was considered and no optimisation algorithm was implemented. To reduce the computational requirements, the utility was evaluated for all possible choices of design, and the design which maximised the utility was chosen. It remains an open question as to how continuous design spaces can be dealt with efficiently in this context.

Should I mention James' new paper here, or just leave it as it is in Section 6?

6 Directions for Future Research

We believe the future of Bayesian experimental design lies in: (1) developing and implementing fast methods for approximating the posterior distribution for use in Bayesian utility functions, and fast computation of the Bayesian utility functions, as these are the most computationally intensive components of Bayesian experimental design; and (2) finding solutions to complex Bayesian experimental design problems, such as problems in which the likelihood is intractable or computationally prohibitive to calculate, or problems with a large number of design variables.

6.1 Fast Algorithms for Bayesian Experimental Design

Computational burden is a major obstacle in Bayesian design problems and must be overcome so that designs can be obtained efficiently and in real time, and to broaden the applicability of Bayesian design methodology by making it more accessible to practitioners, scientists and industry.

In Table 1 we provide a summary of the methods which have previously been used to approximate the posteriors for Bayesian utility functions, along with the search algorithms in which they are embedded.

MCMC and importance sampling have been found to be computationally intensive to perform at each iteration of the optimisation algorithm that searches over the space $(\mathbf{d}, \boldsymbol{\theta}, \mathbf{y})$, due to the large number of samples that are required to ensure that the Bayesian utility is well estimated. In particular, importance sampling from the prior performs poorly when large amounts of data are involved due to the large number of prior simulations that are required to achieve a reasonable ESS (Ryan et al. [2014a]).

Laplace approximations and numerical quadrature have been found to be fast alternatives for approximating the posterior distribution in Bayesian design, and can be used when large amounts of data are involved (e.g., Ryan et al. [2014a]), but rely on the assumption that the posterior distribution follows a multivariate normal distribution and also suffer from the curse of dimensionality. Laplace approximations (to the posterior) have also been used to form the importance distribution for importance sampling (Ryan et al. [2014a]) and can be used when large amounts of data are involved in the design problem and correct for some non-normality that is not accommodated by the Laplace approximation.

Drovandi and Pettitt (2013) and Hainy et al. (2013) have explored the use of ABC rejection (see Beaumont et al. [2002]) within an MCMC framework to approximate the posterior distributions for Bayesian utility functions for design problems in which the likelihood function is intractable. Further use of likelihood-free methods for posterior distribution approximation should be explored in the experimental design context.

A few studies have investigated the use of SMC for approximating the necessary quantities for Bayesian utility functions (e.g., Drovandi et al. [2013]), but its use has been limited. Future studies should focus on extending previous approaches to allow for more complicated design problems.

Search framework	Algorithm	Method for approx. posterior	Example(s)
Static designs			
MCMC		Laplace approximation	Ryan et al. [2014a]
MCMC		Importance sampling	Cook et al. [2008], Ryan et al. [2014a], **?**
MCMC		ABC	Drovandi and Pettitt [2013], Hainy et al. [2013]
MCMC		MCMC	Clyde et al. [1996]
Monte Carlo		MCMC	Han and Chaloner [2004]
SMC		Importance sampling	Amzal et al. [2006]
SPSA and NMNS		Polynomial chaos approximations and nested Monte Carlo integration	Huan and Marzouk [2013]
RM stochastic approximation		Nested Monte Carlo integration	Huan and Marzouk [2012]
SAA-BFGS		Nested Monte Carlo integration	Huan and Marzouk [2012]
Sequential designs			
Discrete search		Laplace approximation	Lewi et al. [2009]
SMC		Numerical quadrature	Cavagnaro et al. [2010]
Discrete search		SMC / importance sampling	Drovandi et al. [2013]
MCMC		Importance sampling	Stroud et al. [2001], McGree et al. [2012c]
Monte Carlo		MCMC	Wakefield [1994], Palmer and Müller [1998]

Table 1: Summary of methods used to approximate the posterior distributions for Bayesian utility function estimation and for optimisation over $(\mathbf{d}, \boldsymbol{\theta}, \mathbf{y})$.

Overcoming computational burden may be achieved through algorithmic developments and the exploitation of current parallel computing technology (such as graphics processing units or GPUs). Indeed, new parallel architectures are becoming increasingly available to individual researchers, and will have a significant impact on Bayesian experimental design. In order to take advantage of this increased power, computational problems and approaches should be adapted from the current serial processing paradigm to one that optimises algorithms for parallel processing. McGree et al. (2014) used GPUs to overcome the computational burden of searching for optimal sequential Bayesian designs for mixed effects models. Their results demonstrated significant improvements in computational speed over Matlab and C implemented code.

6.2 Finding Optimal Designs for Complex Models

The future of Bayesian experimental design also lies in solving complex or nonstandard problems, such as problems in which the likelihood is intractable or computationally prohibitive to evaluate, problems where the observed data likelihood cannot be evaluated analytically, or problems with a large number of design points. Whilst sophisticated inference techniques are available for Bayesian data analysis for complex data models, corresponding methodology for deriving Bayesian experimental designs is severely lacking, and it is important that the methods for inference are complemented with appropriate experimental design methodologies that enable more informative data to be collected in a more timely manner. Use of parallel computing technology may be required to ease the computational burden of finding optimal Bayesian experimental designs for complex models (such as mixed effects models).

Fully Bayesian experimental designs for nonlinear mixed effects models are largely unexplored. Most of the current work has focused on evaluating Bayesian utility functions for a fixed set of discrete designs (e.g., Han and Chaloner [2004], Palmer and Müller [1998]) and selecting the design that produces the highest utility value (i.e., no search over a continuous design space is performed). Ryan et al. (2014b) extend this by searching over a continuous design space to determine (near) optimal sampling times for a horse population pharmacokinetic study. Kim et al. (2013) find optimal sequential designs for population studies. McGree et al. (2014) have recently conducted work on using SMC algorithms (Chopin [2002]) to search for optimal designs for mixed effects models in the presence of parameter and model uncertainty. The main difficulty in finding solutions to experimental design problems in which the data is modelled by mixed effects models is in obtaining good approximations to the posterior for the fixed effects parameters. This is easier if the number of random effects is small and software such as INLA might be useful in this context.

6.3 Finding Optimal Designs for a Large Number of Design Variables

Better search algorithms are also required to find static designs. Many of the search algorithms for obtaining optimal designs (e.g., Müller [1999], Amzal et al. [2006]) are restricted to a small number of design variables (≤ 4), as these algorithms are computationally prohibitive for a large number of design variables (e.g., Bielza et al. [1999], Müller [1999], Stroud et al. [2001], Cook et al. [2008]). MCMC algorithms are good at estimating the marginal distribution of random variables, but experimental design requires the joint distribution, and in particular the joint mode of the design variables, which is very difficult to find and estimate in high dimensions.

Ryan et al. (2014c) propose the use of lower dimensional parameterisations or projections to enable near optimal designs to be found for problems that require a large number of design points. The lower dimensional parameterisations consist of a few design variables, which are optimised, and are then input into various functions to generate multiple design points. This was found to have substantial computational savings, and it was much easier to obtain the multivariate mode for a few design variables than for a large number of design variables. However, designs found using this method are not optimal but *near* optimal, which is a compromise of the computational savings achieved. How close they are to optimal is difficult to investigate. The approach is only useful for design variables (e.g., sampling times/locations) that require multiple measures to be taken at specific points that are separated from one another in the design space. This approach does not overcome the problem of having a large number of different types of design variables (e.g., temperatures, pressures), and further research needs to be conducted for solving this design problem.

7 Conclusion

Bayesian experimental design is a fast growing area of research with many exciting recent developments. The Bayesian approach to experimental design offers many advantages over frequentist approaches, the most notable of which is the ability to optimise design criteria that are functions of the posterior distribution and can easily be tailored to the experimenters' design objectives. Bayesian frameworks also provide a formal approach for incorporating parameter uncertainties and prior information into the design process via prior distributions, and provide a unified approach for joining these quantities with the model and design criterion. The Bayesian approach solves sequential design problems in a principled way, updating a prior to become a posterior as new data are observed. The prior information is not "thrown away" in fully Bayesian experimental design, as it is in pseudo-Bayesian design, but the downfall is that Bayesian design is a harder computational problem.

Whilst several review papers on Bayesian experimental design have been written, there is a lack of recent Bayesian experimental design papers that reflect the computational advancements that have occurred in recent times. In this article we have reviewed the computational methods that have been used to approximate the

posterior distribution for Bayesian utility functions, along with methods for calculating the Bayesian utility functions (once the posterior has been approximated) and the search algorithms that have been used for finding the optimal designs. We have also highlighted some numerical methods and stochastic algorithms that have previously been used to perform Bayesian inference, but have not been used in the design context, and may provide fast alternatives for finding Bayesian designs.

It is our opinion that the future of Bayesian experimental design lies in the development and implementation of rapid methods for approximating the Bayesian utility functions, since this is the most computationally intensive component of the Bayesian experimental design process. We also believe that the future of Bayesian experimental design lies in finding solutions to complex or nonstandard design problems, such as problems in which the likelihood is intractable or computationally prohibitive to evaluate, problems where the observed data likelihood cannot be evaluated analytically, or problems with a large number of design points or design variables. Solutions to these difficult problems can only be achieved through algorithmic developments and the exploitation of current parallel computing technology.

Acknowledgments

E.G. Ryan was supported by an APA(I) Scholarship which came from an ARC Linkage Grant with Roche Palo Alto (LP0991602). The work of A.N. Pettitt was supported by an ARC Discovery Project (DP110100159), and the work of J.M. McGree was supported by an ARC Discovery Project (DP120100269).

References

B. Amzal, F. Bois, E. Parent, and C. P. Robert. Bayesian-optimal design via interacting particle systems. *Journal of the American Statistical Association*, 101(474):773–785, 2006.

A. C. Atkinson. Dt-optimum designs for model discrimination and parameter estimation. *Journal of Statistical Planning and Inference*, 138:56–64, 2008.

A. C. Atkinson and A. N. Donev. *Optimum Experimental Designs*. Oxford University Press, New York, 1992.

M. A. Beaumont, W. Zhang, and D. J. Balding. Approximate Bayesian computation in population genetics. *Genetics*, 162(4):2025–2035, 2002.

T. Bengtsson, P. Bickel, and B. Li. Curse-of-dimensionality revisited: Collapse of the particle filter in very large scale systems. In *Probability and Statistics: Essays in Honor of David A. Freedman*. Institute of Mathematical Statistics, 2008.

J. O. Berger. *Statistical Decision Theory and Bayesian Analysis*. Springer-Verlag, New York, 2nd edition, 1985.

J. M. Bernardo. Expected information as expected utility. *Annals of Statistics*, 7(3):686–690, 1979.

J. M. Bernardo and A. F. M. Smith. *Bayesian Theory*. John Wiley & Sons, 2nd edition, 2000.

D. Berry, P. Müller, A. Grieve, M. Smith, T. Parke, R. Blazek, N. Mitchard, and M. Krams. *Case Studies in Bayesian Statistics*, chapter Adaptive bayesian designs for dose-ranging drug trials. Springer, 2000.

D. A. Berry and C.-H. Ho. One-sided sequential stopping boundaries for clinical trials: A decision-theoretic approach. *Biometrics*, 44:219–227, 1988.

C. Bielza, P. Müller, and D. R. Insua. Decision analysis by augmented probability simulation. *Management Science*, 45(7):995–1007, 1999.

B. Bornkamp, F. Bretz, H. Dette, and J. Pinheiro. Response-adaptive dose-finding under model uncertainty. *Annals of Applied Statistics*, 5(2B):1611–1631, 2011.

D. M. Borth. A total entropy criterion for the dual problem of model discrimination and parameter estimation. *Journal of the Royal Statistical Society: Series B (Methodological)*, 37:77–87, 1975.

G. E. P. Box and W. J. Hill. Discrimination among mechanistic models. *Technometrics*, 9:57–71, 1967.

A. E. Brockwell and J. B. Kadane. A gridding method for Bayesian sequential decision problems. *Journal of Computational and Graphical Statistics*, 12(3):566–584, 2003.

B. Carlin, J. Kadane, and A. Gelfand. Approaches for optimal sequential decision analysis in clinical trials. *Biometrics*, 54(3):964–975, 1998.

D. R. Cavagnaro, J. I. Myung, M. A. Pitt, and J. V. Kujala. Adaptive design optimization: A mutual information-based approach to model discrimination in cognitive science. *Neural Computation*, 22(4):887–905, 2010.

K. Chaloner. Optimal bayesian experimental designs for linear models. *Annals of Statistics*, 12:283–300, 1984.

K. Chaloner. An approach to design for generalised linear models. In *Proceedings of the Workshop on Model-oriented data analysis, Wartburg. Lecture Notes in Economics and Mathematical Systems.*, Berlin, 1987. Springer.

K. Chaloner and I. Verdinelli. Bayesian experimental design: A review. *Statistical Science*, 10:273–304, 1995.

N. Chopin. A sequential particle filter method for static models. *Biometrika*, 89(3):539–552, 2002.

J. Christen, P. Müller, K. Wathen, and J. Wolf. Bayesian randomized clinical trials: A decision-theoretic sequential design. *Canadian Journal of Statistics*, 32(4):387–402, 2004.

M. Clyde and K. Chaloner. The equivalence of constrained and weighted designs in multiple objective design problems. *Journal of the American Statistical Association*, 91:1236–1244, 1996.

M. A. Clyde, P. Müller, and G. Parmigiani. Exploring expected utility surfaces by Markov Chains. Technical report, Duke University, 1996.

A. Cook, G. Gibson, and C. Gilligan. Optimal observation times in experimental epidemic processes. *Biometrics*, 64(3):860–868, 2008.

D. D'Argenio. Incorporating prior parameter uncertainty in the design of sampling schedules for pharmacokinetic parameter estimation experiments. *Mathematical Biosciences*, 99(1):105–118, 1990.

A. DasGupta, S. Mukhopadhyay, and W. Studden. Compromise designs in heteroscedastic linear models. *Journal of Statistical Planning and Inference*, 32:363–384, 1992.

M. H. DeGroot. *Optimal Statistical Decisions*. McGrawHill, New York, 1970.

P. Del Moral, A. Doucet, and A. Jasra. Sequential Monte Carlo samplers. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(3):411–436, 2006.

H. Dette. A generalization of D-and D1-optimal designs in polynomial regression. *The Annals of Statistics*, 18: 1784–1804, 1990.

P. Diggle and S. Lophaven. Bayesian geostatistical design. *Scandinavian Journal of Statistics*, 33(1):53–64, 2006.

C. Drovandi, J. McGree, and A. Pettitt. A sequential Monte Carlo algorithm to incorporate model uncertainty in Bayesian sequential design. *Journal of Computational and Graphical Statistics*, 23(1):3–24, 2014.

C. C. Drovandi and A. N. Pettitt. Estimation of parameters for macroparasite population evolution using approximate Bayesian computation. *Biometrics*, 67(1):225–233, 2011.

C. C. Drovandi and A. N. Pettitt. Bayesian experimental design for models with intractable likelihoods. *Biometrics*, 69(4):937–948, 2013.

C. C. Drovandi, A. N. Pettitt, and M. J. Faddy. Approximate Bayesian computation using indirect inference. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 60(3):503–524, 2011.

C. C. Drovandi, J. M. McGree, and A. N. Pettitt. Sequential Monte Carlo for Bayesian sequential design. *Computational Statistics and Data Analysis*, 57(1):320 – 335, 2013.

S. El-Krunz and W. Studden. Bayesian optimal designs for linear regression models. *Annals of Statistics*, 19: 2183–2208, 1991.

R. Etzione and J. B. Kadane. Optimal experimental design for another's analysis. *Journal of the American Statistical Association*, 88:1401–1411, 1993.

V. Federov and P. Hackl. *Model-oriented Design of Experiments*. Springer-Verlag, Berlin, 1997.

V. V. Fedorov. *Theory of Optimal Experiments*. Academic Press, New York, 1972.

J. Geweke. Bayesian inference in Econometric models using Monte Carlo integration. *Econometrica*, 57(6): 1317–1339, 1989.

M. Hainy, W. Müller, and H. Wagner. Likelihood-free simulation-based optimal design. Technical report, Johannes Kepler University, Linz, 2013.

C. Han and K. Chaloner. Bayesian experimental design for nonlinear mixed-effects models with application to HIV dynamics. *Biometrics*, 60:25–33, 2004.

W. Hill, W. Hunter, and D. Wichern. A joint design criterion for the dual problem of model discrimination and parameter estimation. *Technometrics*, 10(1):145–160, 1968.

W. J. Hill. A review of experimental design procedures for regression model discrimination. *Technometerics*, 20: 15–21, 1978.

Y. C. Ho and X. Cao. Perturbation analysis and optimization of queueing networks. *Journal of Optimization Theory and Applications*, 40:559–582, 1983.

X. Huan and Y. M. Marzouk. Gradient-based stochastic optimization methods in bayesian experimental design. Technical report, Massachusetts Institute of Technology, Cambridge, 2012.

X. Huan and Y. M. Marzouk. Simulation-based optimal Bayesian experimental design for nonlinear systems. *Journal of Computational Physics*, 232(1):288–317, 2013.

J. B. Kadane. *Bayesian methods and ethics in clinical trial design*. John Wiley & Sons, 1996.

J. Kiefer. Optimum experimental designs. *Journal of the Royal Statistical Society. Series B*, 21:272–304, 1959.

J. Kiefer. Optimum designs in regression problems. II. *Annals of Mathematical Statistics*, 32(2):298–325, 1961.

J. Kiefer. General equivalence theory for optimum designs (approximate theory). *Annals of Statistics*, 2(5): 849–1063, 1974.

J. Kiefer and J. Wolfowitz. Optimum designs on regression problems. *Annals of Mathematical Statistics*, 30: 271–94, 1959.

J. Kiefer and J. Wolfowitz. The equivalence of two extremum problems. *Canadian Journal of Mathematics*, 14: 363–366, 1960.

J. Kiefer and J. Wolfowitz. On a theorem of Hoel and Levine on extrapolation designs. *Annals of Mathematical Statistics*, 36(6):1627–1655, 1965.

P. Kinas. Bayesian fishery stock assessment and decision making using adaptive importance sampling. *Canadian Journal of Fisheries and Aquatic Sciences*, 53:414–423, 1996.

H. Kück, N. de Freitas, and A. Doucet. Smc samplers for Bayesian optimal nonlinear design. Technical report, University of British Columbia, 2006.

J. V. Kujala and T. J. Lukka. Bayesian adaptive estimation: The next dimension. *Journal of Mathematical Psychology*, 50(4):369–389, 2006.

S. Kullback and R. A. Leibler. On information and sufficiency. *The Annals of Mathematical Statistics*, 22(1): 79–86, 1951.

L. Kuo, R. Soyer, and F. Wang. *Bayesian Statistics VI*, chapter Optimal design for quantal bioassay via Monte Carlo methods, pages 795–802. Oxford University Press, New York, 1999.

J. Lewi, R. Butera, and L. Paninski. Sequential optimal design of neurophysiology experiments. *Neural Computation*, 21:619–687, 2009.

R. Lewis and D. A. Berry. Group sequential clinical trials: A classical evaluation of Bayesian decision-theoretic designs. *Journal of the American Statistical Association*, 89:1528–1534, 1994.

J. Liepe, S. Filippi, M. Komorowski, and M. P. H. Stumpf. Maximising the information content of experiments in systems biology. *PLoS Computational Biology*, 9(1):e1002888. doi:10.1371/journal.pcbi.1002888, 2013.

D. Lindley. On a measure of the information provided by an experiment. *Annals of Mathematical Statistics*, 27: 986–1005, 1956.

D. Lindley. The choice of variables in multiple regression. *Journal of the Royal Statistical Society Series B*, 30: 31–53, 1968.

D. Lindley. *Bayesian Statistics - A Review*. SIAM, Philadelphia, 1972.

G. Liu, W. F. Rosenberger, and L. M. Haines. Sequential designs for ordinal phase I clinical trials. *Biometrical Journal*, 51(2):335–347, 2009.

J. Loeppky, L. Moore, and B. Williams. Batch sequential designs for computer experiments. *Journal of Statistical Planning and Inference*, 140:1452–1464, 2010.

Q. Long, M. Scavino, R. Tempone, and S. Wang. Fast estimation of expected information gains for Bayesian experimental designs based on Laplace approximations. *Computer Methods in Applied Mechanics and Engineering*, 259:24–39, 2013.

J. López-Fidalgo, C. Tommani, and P. Trandafir. An optimal experimental design criterion for discriminating between non-normal models. *Journal of the Royal Statistical Society:Series B (Statistical Methodology)*, 69: 231–242, 2007.

T. Loredo. Bayesian adaptive exploration. In *Bayesian Inference and Maximum Entropy Methods in Science and Engineering: 23rd International Workshop on Bayesian Inference and Maximum Entropy Methods in Science and Engineering*, pages 330–346, 2004.

J. Maryak and D. Chin. Global random optimization by simultaneous perturbation stochastic approximation. *Johns Hopkins APL Technical Digest*, 25(2):91–100, 2004.

T. A. Masoumi, S.and Duever and P. M. Reilly. Sequential Markov chain Monte Carlo (MCMC) model discrimination. *The Canadian Journal of Chemical Engineering*, 91(5):862–869, 2013.

J. McGree, C. C. Drovandi, and A. N. Pettitt. A sequential Monte Carlo approach to derive sampling times and windows for population pharmacokinetic studies. *Journal of Pharmacokinetics and Pharmacodynamics*, 39(5): 519–526, 2012a.

J. McGree, C. C. Drovandi, and A. N. Pettitt. A sequential Monte Carlo approach to the sequential design for discriminating between rival continuous data models. Technical report, Queensland University of Technology, 2012b.

J. McGree, C. C. Drovandi, H. Thompson, J. Eccleston, S. Duffull, K. Mengersen, A. N. Pettitt, and T. Goggin. Adaptive Bayesian compound designs for dose finding studies. *Journal of Statistical Planning and Inference*, 142(6):1480–1492, 2012c.

P. Müller. Simulation-based optimal design. *Bayesian Statistics*, 6:459–474, 1999.

P. Müller and G. Parmigiani. Optimal design via curve fitting of monte carlo experiments. *Journal of the American Sta*, 90(432):1322–1330, 1995.

P. Müller, B. Sansó, and M. De Iorio. Optimal Bayesian design by inhomogeneous Markov chain simulation. *Journal of the American Statistical Association*, 99(467):788–798, 2004.

P. Müller, D. A. Berry, A. P. Grieve, and M. Krams. A Bayesian decision-theoretic dose-finding trial. *Decision Analysis*, 3(4):197–207, Dec. 2006.

J. A. Nelder and R. Mead. A simplex method for function minimization. *The Computer Journal*, 7(4):308–313, 1965.

S. H. Ng and S. E. Chick. Design of follow-up experiments for improving model discrimination and parameter estimation. *Naval Research Logistics*, 51:1129–1148, 2004.

J. Nocedal and S. J. Wright. *Numerical Optimization*. Springer, 2nd edition, 2006.

J. Ormerod and M. Wand. Explaining variational approximations. *American Statistical Association*, 64(2): 140–153, 2010.

J. Palmer and P. Müller. Bayesian optimal design in population models for haematologic data. *Statistics in Medicine*, 17:1613–1622, 1998.

L. Paninski. Asymptotic theory of information-theoretic experimental design. *Neural Computation*, 17:1480–1507, 2005.

T. Pennanen and M. Koivu. An adaptive importance sampling technique. In H. Niederreiter and D. Talay, editors, *Monte Carlo and Quasi-Monte Carlo Methods 2004*, pages 443–455. Springer Berlin, Heidelberg, 2006.

J. Pilz. *Bayesian estimation and experimental design in linear regregression models (2nd ed)*. Wiley, New York, 1991.

L. Pronzato and E. Walter. Robust experiment design via stochastic approximation. *Mathematical Biosciences*, 75(1):103–120, 1985.

F. Pukelsheim. *Optimal Design of Experiments*. Wiley, New York, 1993.

F. Pukelsheim and B. Torsney. Optimal weights for experimental designs on linearly independent support points. *The Annals of Statistics*, 19(3):1614–1625, 1991.

D. Rios Insua and F. Ruggeri. *Robust Bayesian Analysis*. Springer Verlag, New York, 2000.

H. Robbins and S. Monro. A stochastic approximation method. *The Annals of Mathematical Statistics*, 22(3): 400–407, 1951.

S. Ross. *Introduction to stochastic dynamic programming*. Academic Press, 1983.

D. Rossell and P. Müller. Sequential stopping for high-throughput experiments. *Biostatistics*, 14(1):75–86, 2013.

D. Rossell, P. Müller, and G. L. Rosner. Screening designs for drug development. *Biostatistics*, 8(3):595–608, 2007.

P. Roth. *Design of Experiments for Discrimination Among Rival Models.* PhD thesis, Princeton University, New Jersey, USA., 1965.

H. Rue, S. Martino, and N. Chopin. Approximate Bayesian inference for latent Gaussian models using integrated nested Laplace approximations (with discussion). *Journal of the Royal Statistical Society, Series B*, 71(2): 319–392., 2009.

E. Ryan, C. C. Drovandi, and A. N. Pettitt. Fully Bayesian experimental design for pharmacokinetic studies. Technical report, Queensland University of Techonology, 2014a.

E. Ryan, C. C. Drovandi, and A. N. Pettitt. Simulation-based fully bayesian experimental design for mixed effects models. Technical report, Queensland University of Technology, 2014b.

K. Ryan. Estimating expected information gains for experimental designs with application to the random fatigue-limit model. *Journal of Computational and Graphical Statistics*, 12:585–603, 2003.

C. Shannon. A mathematical theory of communication. *Bell System Technical Journal*, 27:379–423, 623–656, 1948.

S. D. Silvey. *Optimal design*. Chapman and Hall, London, 1980.

S. A. Sisson and Y. Fan. *MCMC handbook*, chapter Likelihood-free Markov chain Monte Carlo, pages 313–335. Chapman & Hall., 2011.

A. Solonen, H. Haario, and M. Laine. Simulation-based optimal design using a response variance criterion. *Journal of Computational and Graphical Statistics*, 21(1):234–252, 2012.

J. C. Spall. An overview of the simultaneous perturbation method for efficient optimization. *Johns Hopkins APL Technical Digest*, 19(4):482–492, 1998.

D. Spiegelhalter, L. Freedman, and M. Parmar. Bayesian approaches to randomize trials. In D. Berry and D. Stangl, editors, *Bayesian Biostatistics*, pages 67–108. Dekker, New York, 1996.

J. Stroud, P. Müller, and G. Rosner. Optimal sampling times in population pharmacokinetic studies. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 50(3):345–359, 2001.

Y. Tian and D. Wang. Sequential bayesian design for estimation of EDp. In *The 2nd International Conference on Biomedical Engineering and Informatics, 2009. BMEI'09*, 2009.

B. Toman and J. Gastwirth. Efficienct robust experimental design and estimation using a data-based prior. *Statistical Sinica*, 4:603–615, 1994.

C. Tommasi. Optimal designs for both model discrimination and parameter estimation. *Journal of Statistical Planning and Inference*, 139:4123–4132, 2009.

C. Tsai and K. Chaloner. *Case studies in Bayesian Statistics 5*, chapter Using Prior opinions to examine sample size in a clinical trial: two examples, pages 409–423. Springer-Verlag, New York, 2002.

P. Van Laarhoven and E. Aarts. *Simulated Annealing: Theory and Applications*. Reider, Dordrecht, 1987.

J. Vanlier, C. Tiemann, P. Hilbers, and N. van Riel. Optimal experimental design for model selection in biochemical networks. *BMC Systems Biology*, 8:20, 2014.

I. Verdinelli. *Bayesian Statistics 4*, volume 4, chapter Advances in Bayesian Experimental Design (with Discussion), pages 467–481. Oxford University Press, Oxford, 1992.

I. Verdinelli. Bayesian design for the normal linear model with unknown error variance. *Biometrika*, 87:222–227, 2000.

J. Wakefield. An expected loss approach to the design of dosage regimens via sampling-based methods. *Journal of the Royal Statistical Society. Series D (The Statistician)*, 43(1):13–29, 1994.

T. H. Waterhouse, J. A. Eccleston, and S. B. Duffull. Optimal design criteria for discrimination and estimation in nonlinear models. *Journal of Biopharmaceutical Statistics*, 19:386–402, 2009.

J. Whitehead and H. Brunier. Bayesian decision procedures for dose determining experiments. *Statistics in Medicine*, 14:885–893, 1995.

J. Whitehead and D. Williamson. Bayesian decision procedures based on logisic regression models for dose-finding studies. *Journal of Biopharmaceutical Statistics*, 8:445–467, 1998.

L. Wolfson, J. Kadane, and M. Small. Expected utility as a policy making tool: an environmental health example. In D. Berry and D. Stangl, editors, *Bayesian Biostatistics*, pages 261–277. Dekker, New York, 1996.

J. Zidek, W. Sun, and N. Le. Designing and integrating composite networks for monitoring multivariate Gaussian pollution fields. *Applied Statistics*, 49:63–79, 2000.